doi: 10.1038/nature07331

Supplementary Information for: Genes mirror geography within Europe

John Novembre[*†] Toby Johnson[‡§¶] Katarzyna Bryc[‖] Zoltán Kutalik[‡¶]
Adam R. Boyko[‖] Adam Auton[‖] Amit Indap[‖] Karen S. King[**]
Sven Bergmann[‡¶] Matthew R. Nelson[**] Matthew Stephens[†‡‡]
Carlos D. Bustamante[‖]

August 4, 2008

[*]Department of Ecology and Evolutionary Biology, Interdepartmental Program in Bioinformatics, University of California-Los Angeles, Los Angeles, California 90024, USA.

[†]Department of Human Genetics, University of Chicago, Chicago, IL 60637, USA.

[‡]Department of Medical Genetics, University of Lausanne, Switzerland.

[§]University Institute for Social and Preventative Medecine, Centre Hospitalier Universitaire Vaudois (CHUV), Lausanne, Switzerland.

[¶]Swiss Institute of Bioinformatics, Switzerland.

[‖]Department of Biological Statistics and Computational Biology, Cornell University, Ithaca, NY, USA.

[**]Genetics, GlaxoSmithKline, Research Triangle Park, NC 27709

[‡‡]Department of Statistics, University of Chicago, Chicago, IL 60637, USA.

Supplementary Notes

Identifying departures from the general pattern of genetics mirroring geography

To identify departures between the PC-based map and geography, we identified individuals who are empirical outliers in PC1-PC2 space relative to their geographic positions (Supplementary Figure 2). Individuals identified as outliers are likely to have mis-specified ancestral origins or have recently migrated. More careful inspection of the demographic information for outlier individuals reveals in many cases circumstances that partially explain their at first unexpected genetic positioning (see next section of Supplemental Information: "Notes on outlier individuals").

We applied a similar approach to detect countries who are empirical outliers with respect to the discrepancy between the PC-based map and geography (Supplementary Fig. 1). There is only one obvious outlier, which is Slovakia; however Slovakia is represented in our data set by only one individual, and based on the individual's position in PC1-PC2 space it's possible this outlier may actually have had Italian, rather than Slovakian ancestry. The Russian Federation is less-striking as an outlier, and appears to lie too far "west" genetically, which may be a result of small sample size ($n = 6$) or simply that the Russians sampled here have ancestry from a location further west than the proxy location for Russia (Moscow) would suggest.

We find similar results when we plot Euclidean distances between countries in PC1-PC2 space vs. geographic distances, and find a strong correlation between the two (Supplementary Fig. 5a, $r^2 = 0.68$). Many of the empirical outliers are pairwise comparisons involving either Slovakia (SK) or Russia (RU). In addition, even after excluding Russia and Slovakia (Supplementary Fig. 5), many of the pairwise comparisons with large residuals involve comparisons with countries that have small sample sizes, [e.g., Kosovo (KS), Slovenia (SI), Scotland (Sct), Finland (FI), Cyprus (CY), Yugoslavia (YG), Croatia (HR)]. This suggests that outlier points are simply due to sampling variation, and not strong departures from a general model where PC1-PC2 position is principally determined by geography.

This pattern is supported when we sample bootstrap distributions on the mean PC1-PC2 position for each country by bootstraping over individuals within each country. We note this bootstrap analysis is approximate in that we use PC1 and PC2 values computed from the original sample (i.e., we do not recompute PC1 and PC2 for each bootstrap sample of individuals, nor do we bootstrap over loci).[1] Furthermore for small samples, bootstrapping is biased (it underestimates the sampling variance–an extreme case being that, for countries with a single individual observed, the bootstrap distributions show zero variance) and it produces artifacts in sampling bootstrap distributions (e.g. discontinuous, multimodal distributions; the problems would be worse if we bootstrapped the median rather than mean positions). With these caveats in mind, the bootstrap distributions (Supplementary Fig. 6) convey that mean PC1 and PC2 positions are typically much more poorly estimated for countries from Eastern and Northern Europe. This further clarifies

[1] A complication to fully bootstrapping PCA results is that the sampling with replacement step necessarily replicates individuals in the sample. For large-scale SNP data, PCA is very sensitive to the presence of replicated individuals, and typically separates such individuals from the rest of the sample on their own axis of variation.

how the observation of countries from Eastern and Northern Europe as outliers in Supplementary Figs. 1 and 5 are more likely a function of small sample sizes, then a true biological signal for differentiation.

In conclusion, the plots of country PC1-PC2 position vs. geographic position have no obvious outliers that cannot be explained plausibly by small sample size and/or the pitfalls of assuming a single proxy location for a large country (e.g. Russia). While there may be more subtle signals of unique population history in the data, the absence of empirical outliers from well-sampled countries suggests that the dominant signal in the data is that the genetics of European populations mirrors their geography.

Notes on outlier individuals

Here we list more details on the demographic information for a sample of outlier individuals. The examples show how the rule of using reported grandparental origins as a proxy for genetic ancestry can in rare cases be misleading. Sometimes country of birth, parental origins, or language information are more useful.

- Individual 44556 has 4 Italian grandparents, but was born in France, speaks French, and clusters with French individuals.
- Individual 7147 has 4 Russian grandparents, but was born in Romania, speaks Romanian, and is placed between Switzerland and Romania in the PC1-PC2 plot.
- Individual 43874 has 4 Swiss grandparents, but was born in Italy, speaks Italian, and clusters with Italian individuals.
- Individual 14215 has 4 Swiss grandparents but has parents who are Italian, speaks Italian, and clusters Italian. Israel is the individual's country of birth.
- Individual 34088 has 4 German GPs but was born in Hungary, speaks Hungarian, and interestingly clusters with Italian individuals.
- The cluster of 5 outlier Italian individuals located well "southwest" of Italy, includes 3 individuals with unobserved grandparental origins and the other 2 have all four grandparents from Italy. Three of the five speak Italian, the other two have unobserved language data. Notably one of the individuals is from the LOLIPOP sub-study and the others are from Lausanne - so both studies identified these outlier Italian individuals, making it unlikely to be the result of some artifact that occurred within one of the two sub-studies from which we draw our data.
- Individual 13011 was born in Slovakia but has no observed grandparental or language information.

References

List of Figures

List of Tables

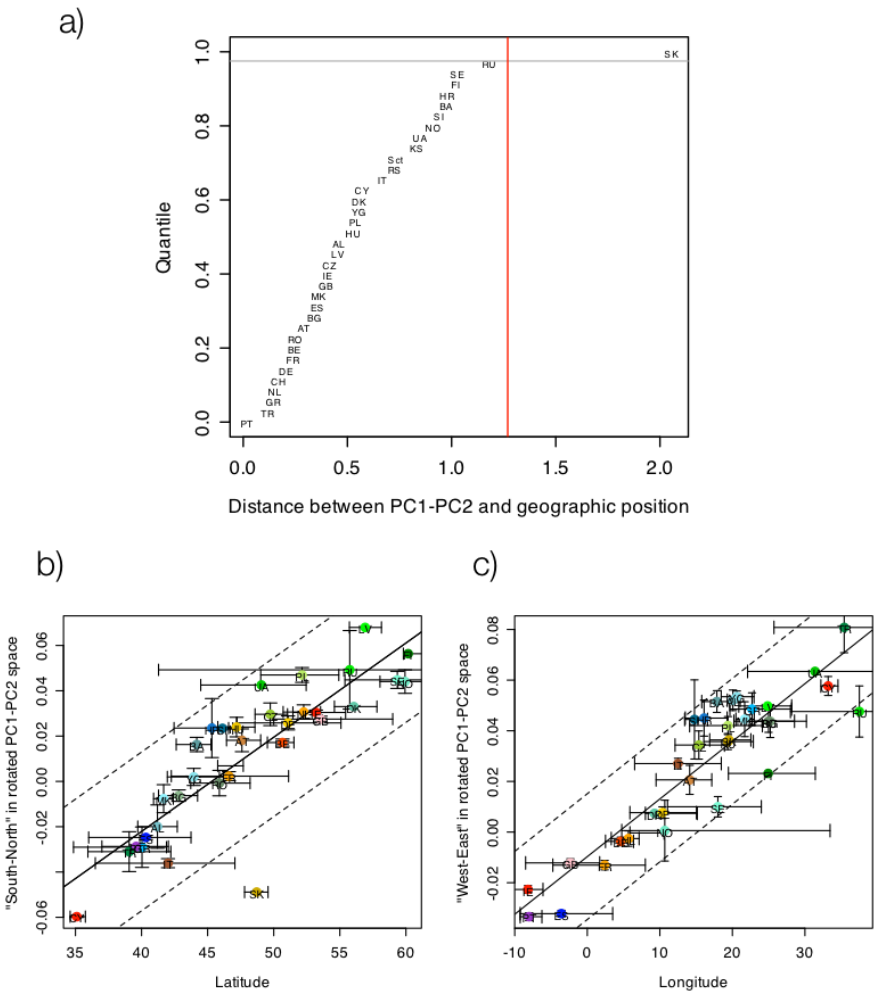

Figure 1: **Detection of outlier countries.** a) Cumulative density of the distances between each country's PC1-PC2 position and their geographic position. To put PC coordinates and geographical coordinates on the same scale, coordinates along each axis are normalized to have mean 0 and standard deviation 1 before computing the distances. An empirical approach is taken whereby outliers are labeled as individuals falling in the upper 2.5% tail of this distance distribution (97.5% quantile denoted by horizontal gray line, corresponding distance threshold marked by vertical red-line). b) Median "south-north" position for each country vs. latitude. c) Median "west-east" position for each country vs. longitude. In (b) and (c) vertical error bars are intervals of +/- 2 standard errors. For samples with 1 individual, error bars are omitted. Horizontal error bars denote the maximal and minimal latitudinal/longitudinal range of the country (n.b., these values are approximate and for Russia we use rough values corresponding to the European portion of Russia). The solid line is based on a regression of position in PC-space vs. geographic space ($r^2 = 0.71$ PC-coordinate vs. latitude; $r^2 = 0.82$ PC-coordinate vs. longitude). Dashed lines denote a region in which 97.5% of all countries fall with respect to the magnitude of the residuals of the regression.

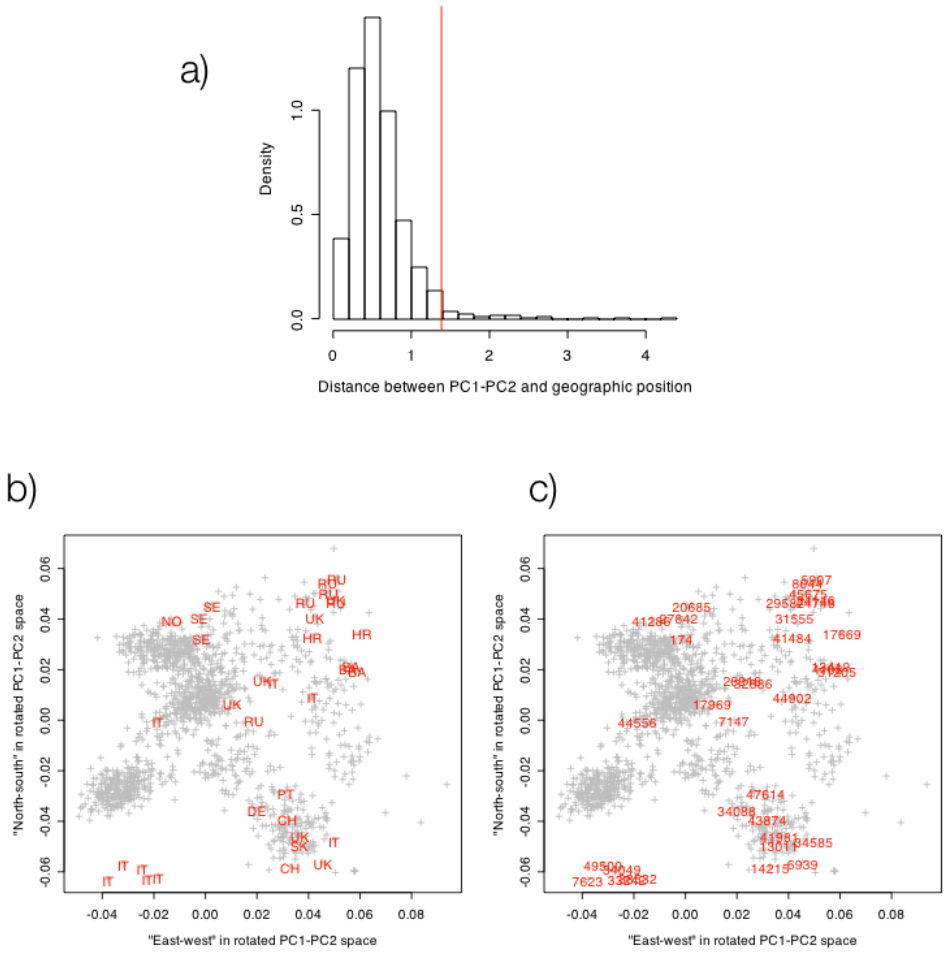

Figure 2: **Detection of outlier individuals.** a) Distribution of the distances between an individual's PC1-PC2 position vs. their geographic position. To put PC coordinates and geographical coordinates on the same scale, coordinates along each axis are normalized to have mean 0 and standard deviation 1 before computing the distances. An empirical approach is taken whereby outliers are labeled as individuals falling in the upper 2.5% tail of this distance distribution (thershold marked by vertical red-line). b) PC1 vs PC2 as in figure 1, with outliers highlighted by ancestry label. c) PC1 vs PC2 as in figure 1, with outliers highlighted by POPRES ID number.

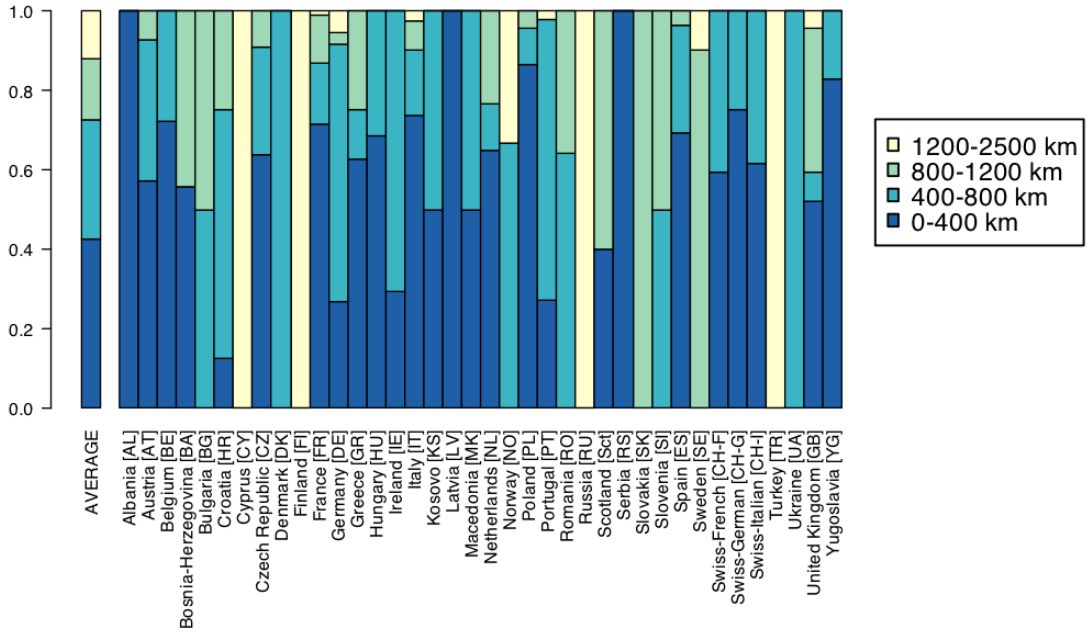

Figure 3: **Prediction accuracy across all populations**. Distances are measured between the population assigned by the discrete assignment method and the origin of the individual determined by grandparental ancestry or country-of-birth. The average column shows the average of the proportions across populations (where each population is given equal weight).

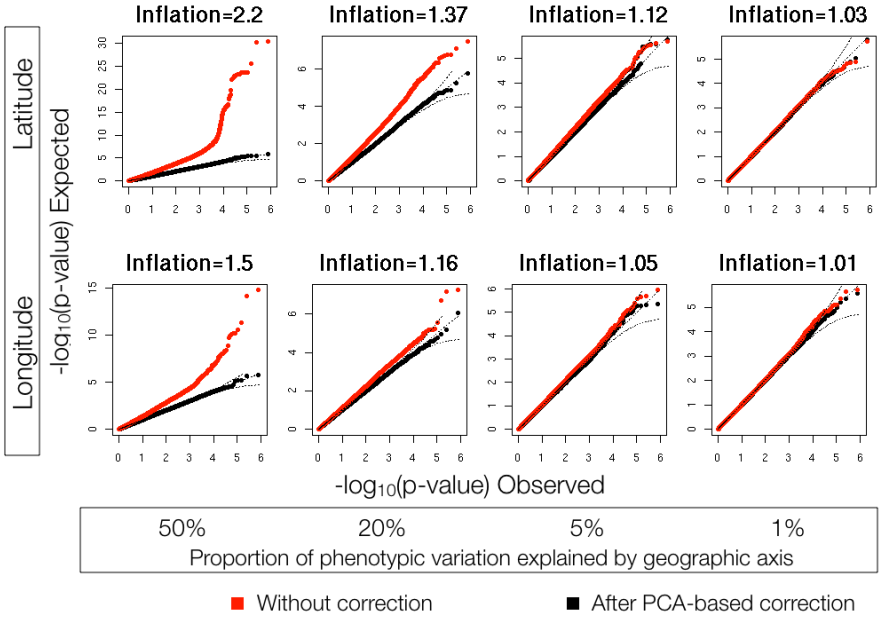

Figure 4: **Quantile-quantile plots for $-\log_{10}$(p-value) distributions from simulated genome-wide association studies where phenotype mean is a function of latitude (top row) or longitude (bottom row).** The phenotype is simulated to have no underlying genetic basis and a mean trait value that depends on latitude or longitude with residual variance about the mean. Each column represents a different percentage of total phenotypic variance explained by latitude or longitude. The inflation statistic (see Methods) takes values of 1 when the observed distribution of p-values matches the expected distribution; values greater than one might reflect an underlying genetic basis to the trait, but in this case the phenotype has no genetic basis, so values greater than one are evidence of spurious association. The dashed lines mark the (0.05,0.5,0.95) quantiles for the observed p-values under the null distribution of no underlying genetic effect.

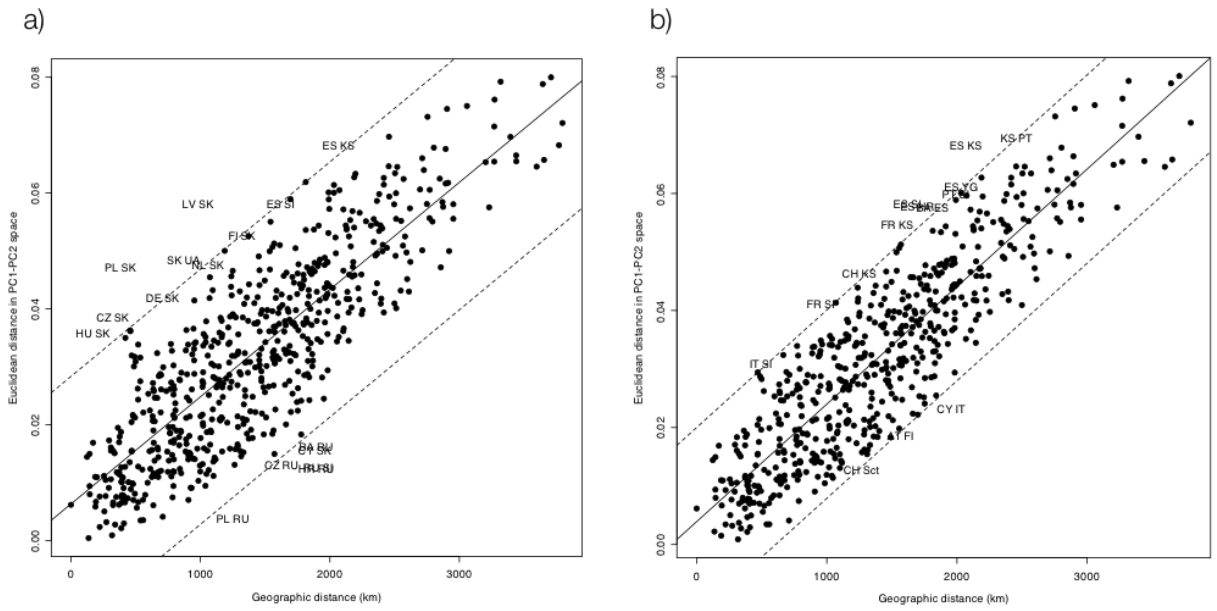

Figure 5: **Distances between countries in PC1-PC2 space vs. distance in geographic space.** Countries positions are taken to be the median PC1-PC2 positions. a) All countries ($r^2 = 0.68$) b) All countries after excluding individuals detected as outliers from Supplementary Figure 1 (which completely excludes Russia and Slovakia as countries) ($r^2 = 0.78$). The solid lines shows the the result of fitting a linear model. The dashed lines indicate the upper and lower limits of residuals falling in the upper 97.5^{th} percentile of residual magnitudes. Paired two-letter country abbreviations mark pairwise distances that fall outside the 97.5^{th} percentile limits.

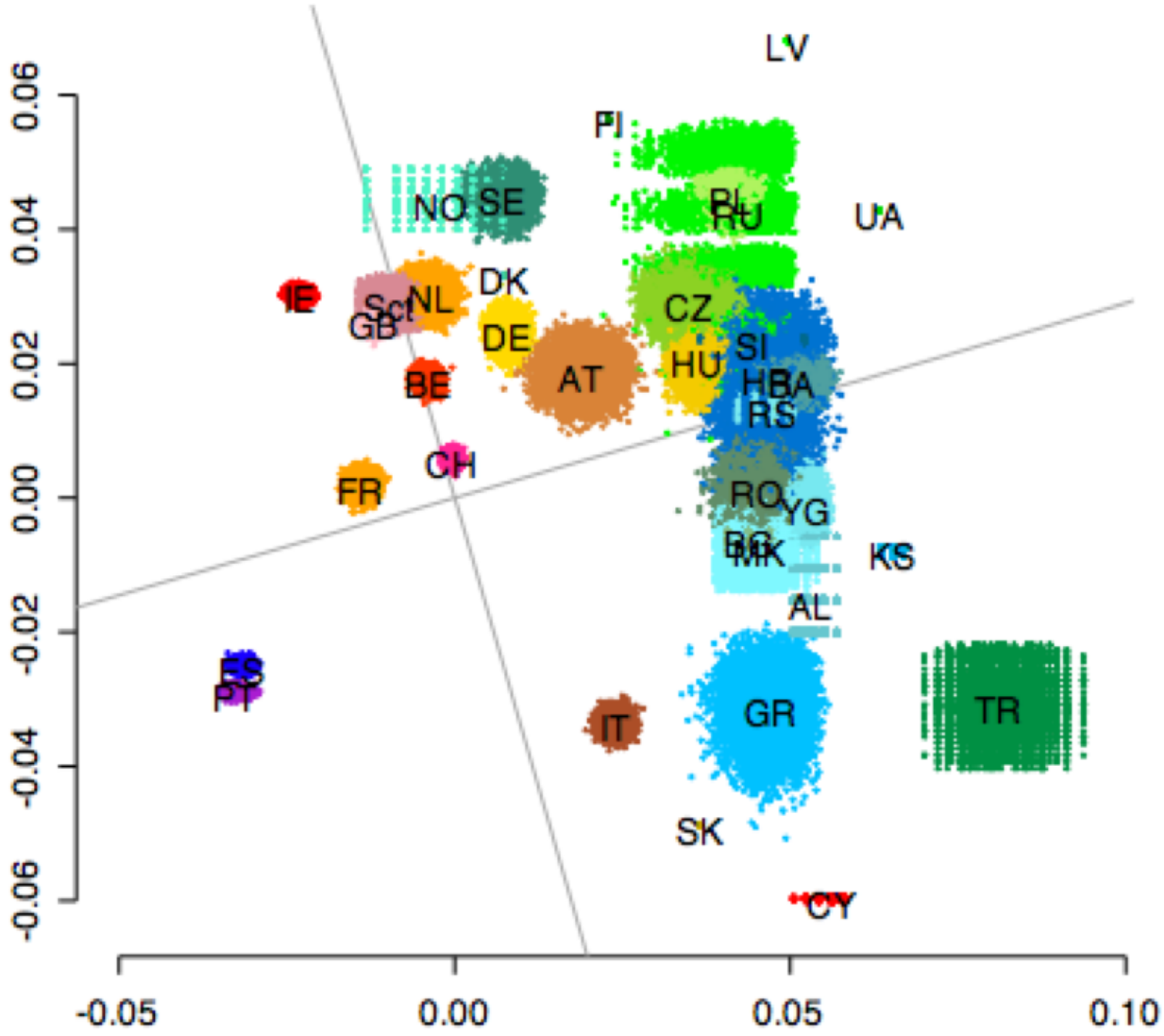

Figure 6: **Mean "north-south" PC-position per country vs. mean "east-west" PC-position per country plotted from 10,000 bootstrapping iterations.**

Number of grandparental origins observed	Number of individuals
0	607
1	0
2	7
3	2
4	771
Total	1387

Table 1: Summary of observed grandparental information for 1387 individuals used in final sample.

Sample size	Stage of analysis
3192	Total individuals of European descent in POPRES Delivery 3
2933	After exclusion of individuals with origins outside of Europe
2409	After exclusion of individuals with mixed grandparental ancestry
2385	After exclusion of putative relateds
2351	After exclusion based on preliminary PCA run
1387	After thinning Swiss-French and UK individuals

Table 2: Summary of number of individuals at each step of sample preparation

.

Geographic Origin	Abbreviation	n	Latitude	Longitude	Group
Italy	IT	219	42	12.5	S
United Kingdom	GB	200	53.5	-2.33	NW
Spain	ES	136	40.3	-3.57	SW
Portugal	PT	128	39.6	-8.02	SW
Swiss-French	CH-F	125	46.2	6.15	W
France	FR	91	46.6	2.39	W
Swiss-German	CH-G	84	47.4	8.55	C
Germany	DE	71	51.1	10.4	C
Ireland	IE	61	53.2	-8.18	NW
Serbia and Montenegro	YG	44	43.9	20.6	SE
Belgium	BE	43	50.7	4.61	W
Poland	PL	22	52.1	19.4	NE
Hungary	HU	19	47.2	19.4	E
Netherlands	NL	17	52.3	5.67	C
Austria	AT	14	47.6	14.1	C
Romania	RO	14	45.9	25	SE
Swiss-Italian	CH-I	13	46	8.95	S
Czech Republic	CZ	11	49.7	15.4	E
Sweden	SE	10	59.4	18	N
Bosnia and Herzegovina	BA	9	44.2	17.9	SE
Croatia	HR	8	45.3	16.1	SE
Greece	GR	8	40	22.7	SE
Russian Federation	RU	6	55.8	37.5	NE
Scotland	Sct	5	56	-3.2	NW
Cyprus	CY	4	35.1	33.2	ESE
Macedonia	MK	4	41.7	21.7	SE
Turkey	TR	4	39.1	35.4	ESE
Albania	AL	3	41.2	20.1	SE
Norway	NO	3	59.9	10.7	N
Bulgaria	BG	2	42.8	25.2	SE
Kosovo	KS	2	42.7	21.1	SE
Slovenia	SI	2	46.1	14.8	SE
Denmark	DK	1	56.1	9.25	N
Finland	FI	1	60.2	24.9	NE
Latvia	LV	1	56.9	24.9	NE
Slovakia	SK	1	48.7	19.5	E
Ukraine	UA	1	49.1	31.4	NE

Table 3: Summary of sample sizes, geographic coordinates used for each origin, and group labels
.

		Discrete Assignment[1]				Continuous[2]		Fixed[3]
Geographic Origin	n	Correct origin[4]	Correct group[4]	50%[5]	90%[5]	50%[5]	90%[5]	
Italy	219	0.74	0.83	0	677	248	576	632
United Kingdom	200	0.5	0.6	284	894	402	755	1931
Spain	136	0.69	0.96	0	498	246	531	2111
Portugal	128	0.27	0.98	498	498	395	759	2599
Swiss-French	125	0.38	0.79	294	417	229	471	893
France	91	0.71	0.84	0	932	293	660	1299
Swiss-German	84	0.31	0.33	294	567	285	482	615
Germany	71	0.27	0.65	534	711	421	821	563
Ireland	61	0.3	1	629	647	451	786	2538
Serbia and Montenegro	44	0.27	0.91	265	521	225	386	816
Belgium	43	0.7	0.91	0	521	241	413	1102
Poland	22	0.77	0.77	0	497	323	487	758
Hungary	19	0.16	0.68	358	531	283	489	582
Netherlands	17	0.41	0.41	212	894	355	553	1064
Austria	14	0.29	0.5	270	615	343	649	0
Romania	14	0	0.86	599	807	631	797	1214
Swiss-Italian	13	0.23	0.54	310	587	374	693	597
Czech Republic	11	0.55	0.64	0	573	243	791	270
Sweden	10	0	0.9	1029	1083	1188	1662	1329
Bosnia and Herzegovina	9	0	0.67	358	807	498	699	552
Croatia	8	0	0.5	466	984	595	876	328
Greece	8	0	0.75	316	1147	488	927	1240
Russian Federation	6	0	0.83	2034	2362	1832	2562	2710
Scotland	5	0	0.4	1044	1054	646	849	2123
Cyprus	4	0	0	1924	2253	2074	2154	2461
Macedonia	4	0	1	312	498	330	463	1046
Turkey	4	0	0	1474	1539	1349	1555	2506
Albania	3	0.33	1	186	192	124	190	946
Norway	3	0	0.67	453	1366	806	1259	1394
Bulgaria	2	0	1	641	788	684	752	1325
Kosovo	2	0	1	290	423	342	381	930
Slovenia	2	0	0.5	824	1068	648	838	174
Denmark	1	0	0	576	576	397	397	1068
Finland	1	0	0	1575	1575	1431	1431	1779
Latvia	1	1	1	0	0	345	345	1537
Slovakia	1	0	0	1056	1056	938	938	608
Ukraine	1	0	0	775	775	561	561	1916
Mean across all populations	37.5	0.24	0.63	537	836	574	809	1231
Mean when n˘6	61.2	0.34	0.73	305	700	398	694	1047

[1] Individuals are assigned to the nearest possible geographic origin based on the latitude and longitude predicted by the linear model (see text)
[2] Individuals are assigned to the latitude and longitude predicted by the linear model
[3] As a reference point, we assess performance if all individuals are assigned to a central point in Europe (here taken to be Austria). Distances to Austria are given in kilometers. [4] Proportion of individuals assigned correctly to the correct geographic origin/group. [5] Quantiles of the distribution of distances in kilometers between assigned location and observed geographic origin.

Table 4: Summary of assignment results using both discrete and continuous assignment methods. Because Europe has many closely spaced countries, assigned locations can be geographically close to the correct origin (per country median error often less than 500km) even though the proportions of individuals assigned correctly to a specific origin/group are low.

Geographic Origin	n	PC1 mean	PC1 SD	PC2 mean	PC2 SD
Italy	219	-0.034	0.0149	0.0234	0.0153
United Kingdom	200	0.0266	0.00937	-0.0114	0.0118
Spain	136	-0.0243	0.00535	-0.0321	0.00782
Portugal	128	-0.0282	0.00415	-0.0332	0.00886
Swiss-French	125	0.00575	0.00428	-0.0028	0.00655
France	91	0.00219	0.00939	-0.0138	0.0089
Swiss-German	84	0.01	0.00549	0.00224	0.00559
Germany	71	0.0247	0.013	0.00852	0.00969
Ireland	61	0.0308	0.00405	-0.0224	0.00594
Serbia and Montenegro	44	-0.00127	0.013	0.0521	0.00727
Belgium	43	0.0177	0.00513	-0.00367	0.00545
Poland	22	0.0442	0.00759	0.042	0.00842
Hungary	19	0.02	0.00944	0.0365	0.00688
Netherlands	17	0.0305	0.00648	-0.00296	0.0063
Austria	14	0.0183	0.00924	0.0194	0.0109
Romania	14	0.000334	0.0105	0.0451	0.00822
Swiss-Italian	13	-0.0173	0.0166	0.00806	0.0125
Czech Republic	11	0.0286	0.00825	0.0354	0.00926
Sweden	10	0.0449	0.00589	0.00825	0.00643
Bosnia and Herzegovina	9	0.0166	0.0047	0.0506	0.00647
Croatia	8	0.0171	0.0184	0.0471	0.01
Greece	8	-0.0321	0.0125	0.0464	0.00928
Russian Federation	6	0.0421	0.021	0.0432	0.0127
Scotland	5	0.0293	0.00373	-0.00902	0.00449
Cyprus	4	-0.0608	0.000288	0.0549	0.00375
Macedonia	4	-0.0082	0.0062	0.0453	0.00687
Turkey	4	-0.0326	0.00874	0.0807	0.01
Albania	3	-0.0164	0.0083	0.0528	0.00348
Norway	3	0.0444	0.00432	-0.000989	0.0105
Bulgaria	2	-0.00704	0.00186	0.0437	0.00453
Kosovo	2	-0.00932	0.000982	0.0651	0.0023
Slovenia	2	0.0226	0.000255	0.0449	0.0111
Denmark	1	0.0329	NA	0.00796	NA
Finland	1	0.056	NA	0.0241	NA
Latvia	1	0.0671	NA	0.051	NA
Slovakia	1	-0.0495	NA	0.0355	NA
Ukraine	1	0.0414	NA	0.0642	NA

Table 5: Mean and standard deviation (SD) of PC1 and PC2 coordinates for each population.